Roadrunner Goes Beyond MPI

Next-Generation Scalable Applications: When MPI-only is not enough

3-5 June 2008

John A. Turner

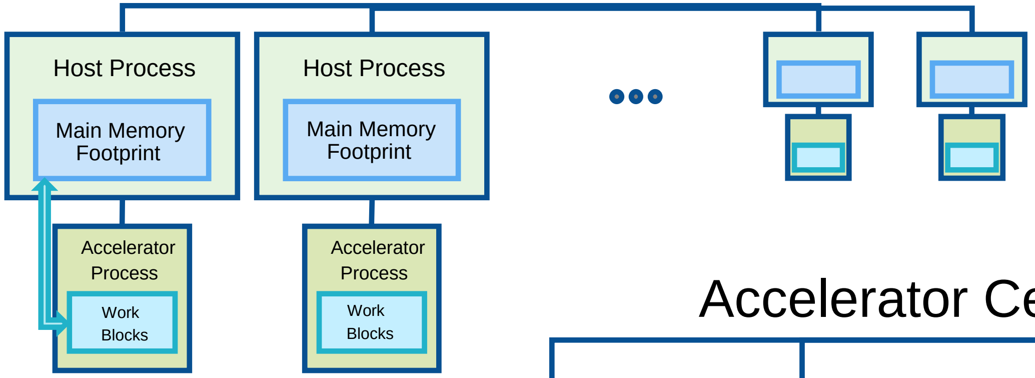
Computational Physics (CCS-2) Group Leader
Computer, Computational, and Statistical Sciences Division (CCS)
Los Alamos National Laboratory (LANL)

How should app developer view Roadrunner?

- **Roadrunner has**
  - ~3200 compute nodes, each with 2 dual-core Opterons
  - ~6400 dual-core Opterons
  - ~13k Opteron cores
  - ~13k Cell processors, each with 8 SPEs
  - ~100k SPEs

Programming Approaches for Roadrunner

Host Centric view



Host Process

Main Memory Footprint

Accelerator Process

Work Blocks

Host Process

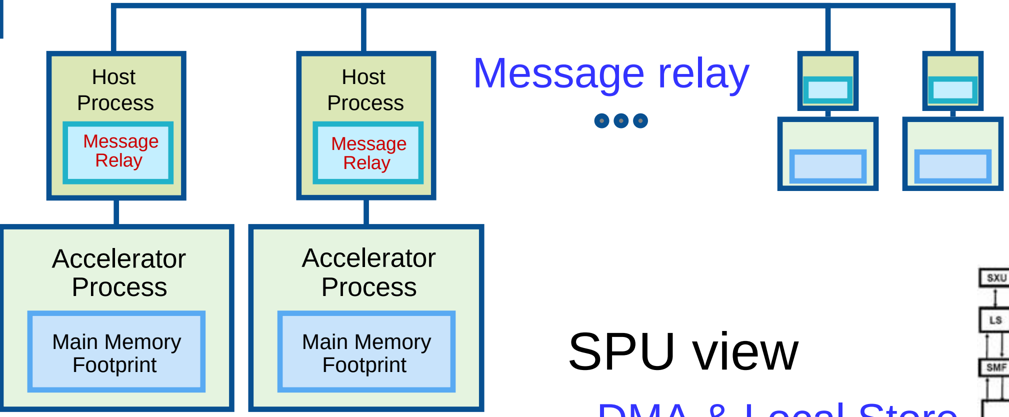
Main Memory Footprint

Accelerator Process

Work Blocks

Function offload

Accelerator Centric view



Host Process

Message Relay

Accelerator Process

Main Memory Footprint

Host Process

Message Relay

Accelerator Process

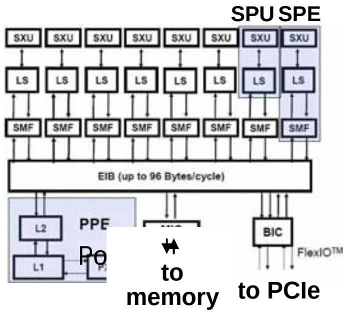
Main Memory Footprint

Message relay

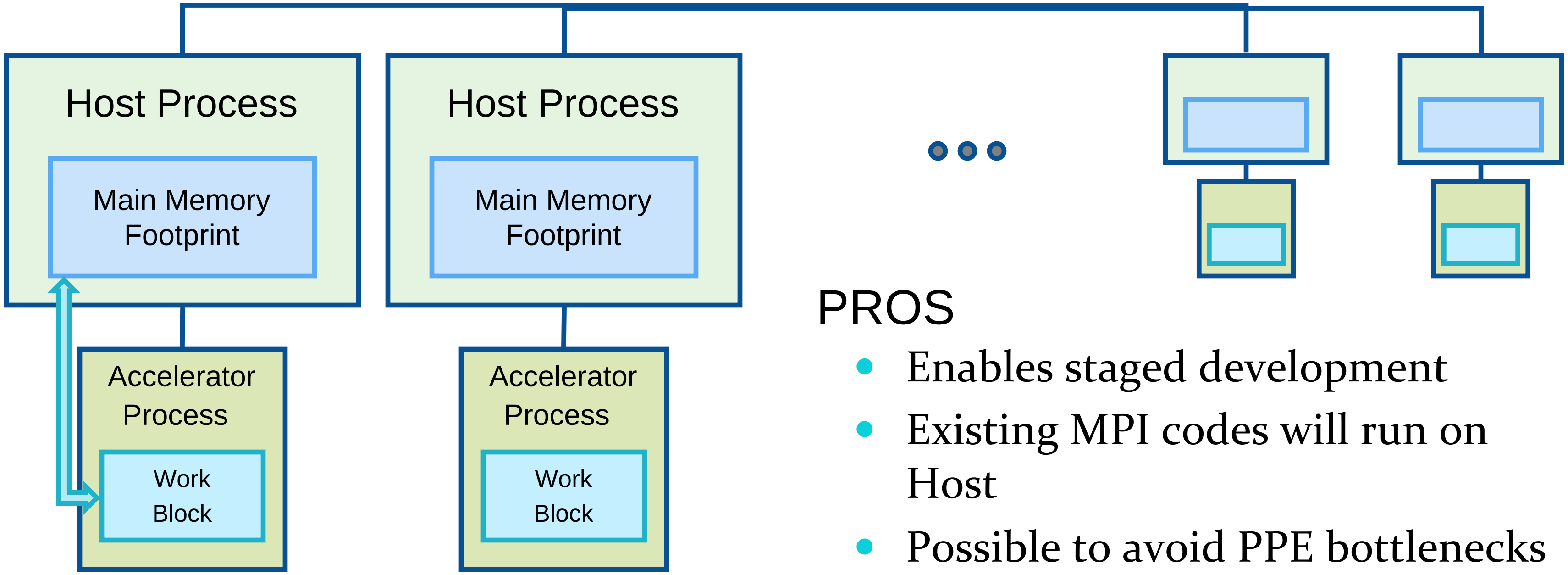
SPU view

DMA & Local Store

SIMD vector



Host-centric model (function offload)



Synchronous or asynchronous function offload to accelerator

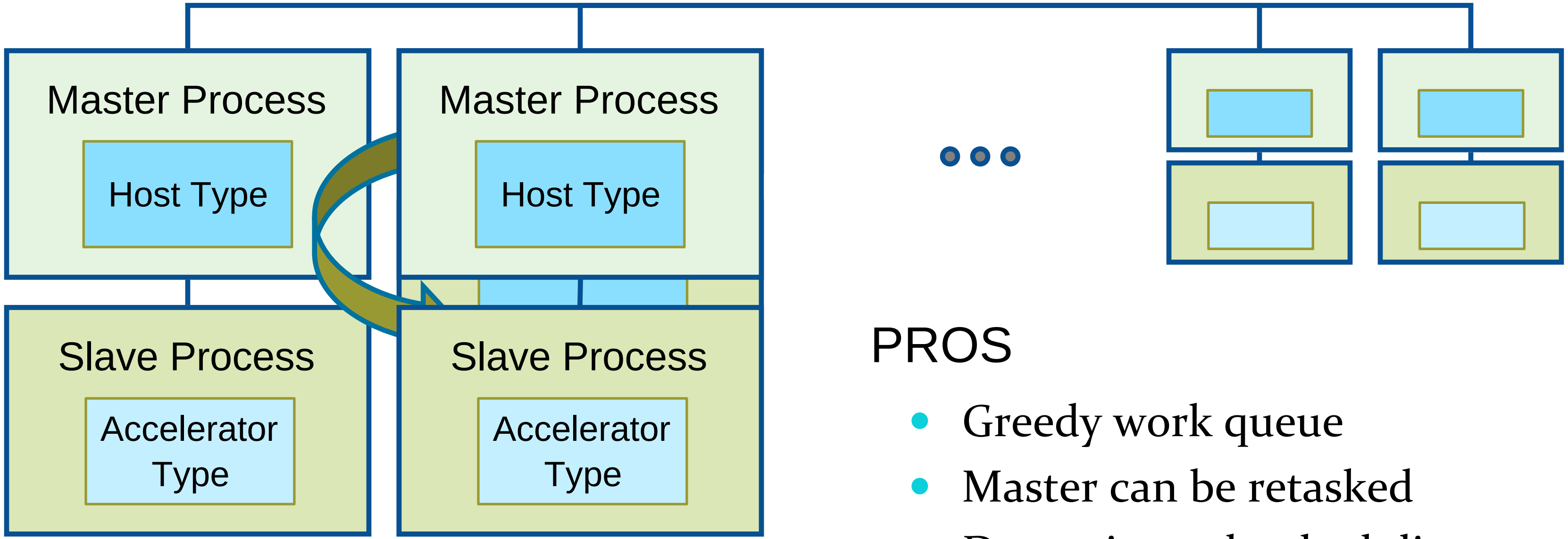
PROS

- Enables staged development
- Existing MPI codes will run on Host
- Possible to avoid PPE bottlenecks

CONS

- Potential data-movement bottleneck

Host-centric model (work stealing)



First step towards true heterogeneous processing

PROS

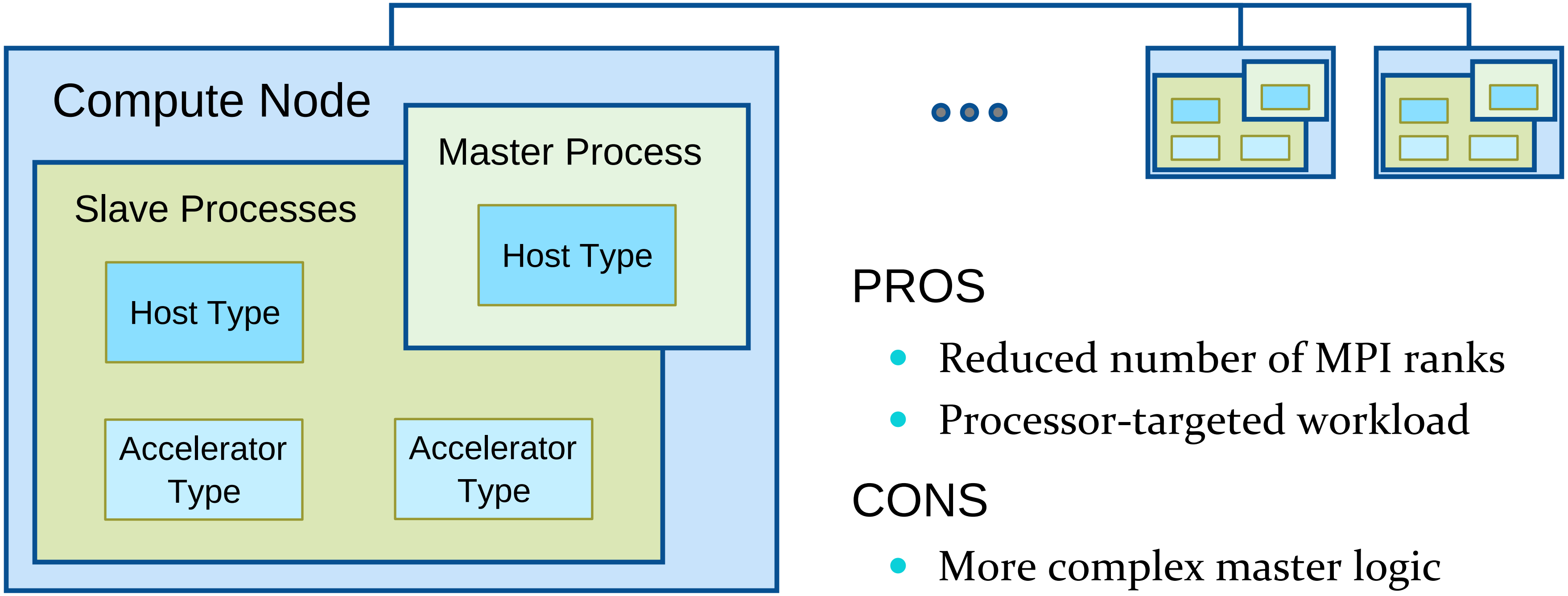
- Greedy work queue
- Master can be retasked
- Dynamic work scheduling

CONS

- More complex master logic

UNCLASSIFIED

Control process model



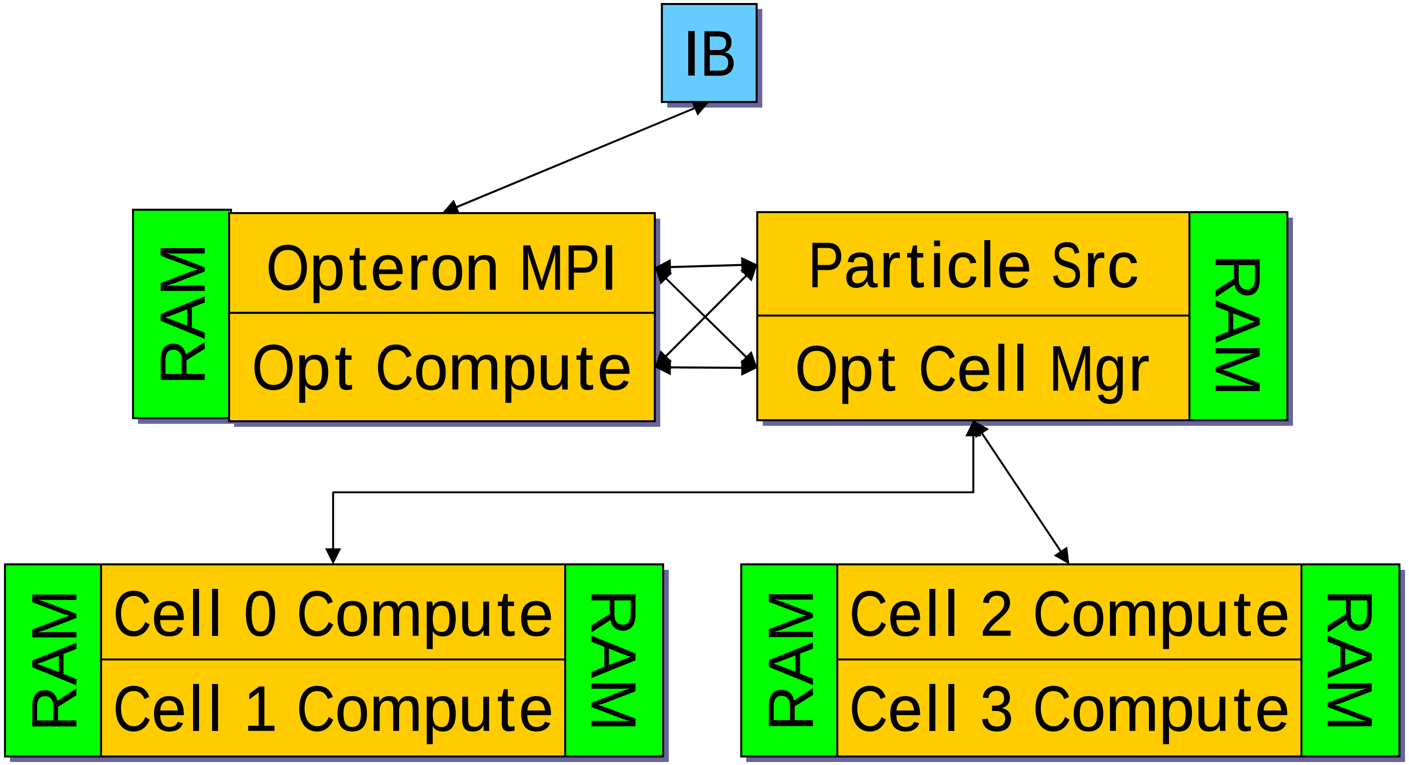
PROS

- Reduced number of MPI ranks
- Processor-targeted workload

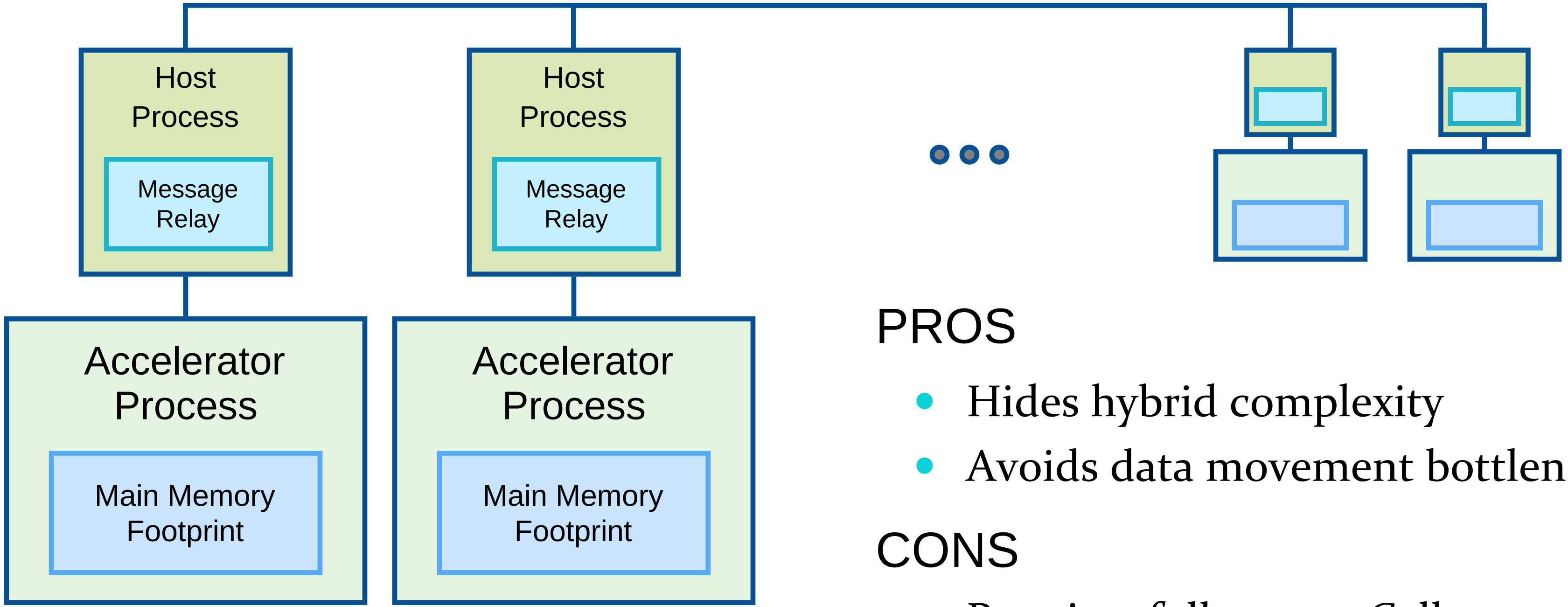
CONS

- More complex master logic
- Software support (DaCS/MPI)

Conceptual control process example for Implicit Monte Carlo (IMC) radiation transport



Accelerator-centric model



MPI traffic relayed through host

PROS

- Hides hybrid complexity
- Avoids data movement bottleneck

CONS

- Requires full port to Cell
- Potential PPE bottleneck

RR hardware is a step to future architectures

- **Future architectures are built on large numbers of thread execution units**
  - RR: ~140K compute threads
  - Sequoia (proposed LLNL machine): O(1.5M) threads
  - Exascale: O(1B) threads
- **Specialized processors likely in commodity market**
- **No clear hardware configuration path:**
  - Hybrid accelerator
    - heterogeneous tree of processors (RR, GPGPU & FPGA cards)
  - Hybrid peers
    - heterogeneous processors on one bus/socket (Torrenza, Intel, AMD, Cray)
  - Many-core/thread
    - homogeneous processors on one socket (BG, all CPU vendors)
- **Some vector capability is certain, but the vector length isn't**

Roadrunner offers applications a spectrum of programming models.

- **Roadrunner has**
  - ~3200 compute nodes, each with 2 dual-core Opterons
  - ~6400 dual-core Opterons
  - ~13k Opteron cores
  - ~13k Cell processors, each with 8 SPEs
  - ~100k SPEs
- **which programming model will provide the best balance of performance, portability, productivity, longevity, etc.?**
  - MPI + threads
    - DaCS + libspe2, DaCS + ALF, hybrid DaCS, hybrid ALF
    - OpenMP, pthreads, TBB, Ct, Cuda, etc.
  - DARPA/HPCS language
    - Chapel, Fortress, X10
  - Partitioned Global Address Space (PGAS) approach
    - GA, UPC, CoArray Fortran

Roadrunner on the web

- http://www.lanl.gov/roadrunner/
- http://en.wikipedia.org/wiki/IBM_Roadrunner

More information on Cell

- **Wikipedia entry on Cell processor**
  - http://en.wikipedia.org/wiki/Cell_processor
- **IBM developerWorks Cell B.E. resource center**
  - http://www-128.ibm.com/developerworks/power/cell/
- **IBM Journal of Research & Development issue devoted to Cell**
  - http://www.research.ibm.com/journal/rd51-5.html
- **IBM developerWorks series on programming the Cell**
  - http://www.ibm.com/developerworks/power/library/pa-linuxps3-1
  - http://www.ibm.com/developerworks/power/library/pa-linuxps3-2
  - http://www.ibm.com/developerworks/power/library/pa-linuxps3-3
- **Power.org Cell Developer Corner (links to tons of info)**
  - http://www.power.org/resources/devcorner/cellcorner/

More information on Cell (cont.)

- **Maximizing the power of the Cell Broadband Engine processor: 25 tips to optimal application performance**
  - http://www.ibm.com/developerworks/library/pa-celltips1/
- **Sony Computer Entertainment US Research and Development**
  - http://www.research.scea.com/
- **MIT course on programming the Playstation 3**
  - http://cag.csail.mit.edu/ps3/index.shtml
- **CellPerformance**
  - http://www.cellperformance.com/
- **Beyond3D.com Cell Forum**
  - http://forum.beyond3d.com/forumdisplay.php?f=57
  - list of Cell resources
    - http://forum.beyond3d.com/showthread.php?t=42626